BIOS 645 Homework #3

1. Examine the plots of residuals vs. predicted values on the second page below. List the letters for the plots that show heteroscedasticity.

Plots B, E, G, and I show heteroscedasticity because the residuals are not uniformly distributed. Rather, the variance of points is higher in one area of the graph compared to others.

1. Examine the qq-plots on the third page below. List the letters for the plots that show non-normality.

Plots A, B, D, E, and F show non-normality because the points on the graph do not all follow a 45-degree line.

1. Lastly, fit a model regressing lung cancer rates onto cigarette sales from the smoke and   
   cancer dataset. When you settle on a final model that you think is a reasonable approximation of the underlying association, report those residual and diagnostic plots, parameter estimates, the hypothesis tests of those estimates, and interpretations.

Regression attempt #1: Lung cancer deaths vs cigarette sales (on page below)

Regression attempt #2: Ln(lung cancer deaths) vs Cigarette sales

Chart, scatter chart

Description automatically generated

Regressing ln(lung cancer deaths) on cigarette sales gave us a qq-plot that appears non-normal, because the center of the line is not at a 45-degree angle.

qq-plot

Residual plot

Regression attempt #3: 1/(lung cancer deaths) vs Cigarette sales

Chart, scatter chart

Description automatically generated

Regressing 1/(lung cancer deaths) on cigarette sales gave us a qq-plot that appears non-normal and a residual plot that looks slightly heteroscedastic.

qq-plot

Residual plot

Diagram, engineering drawing

Description automatically generated

The first, and best, model I tried was regressing y=lung cancer deaths onto x=cigarette sales.

In this model, we can see that the qq-plots for lung cancer deaths almost completely follows a 45-degree line, except for outliers. The residual plot is skewed slightly but the scatter plot is fairly homoscedastic, not following a funnel shape.

Graphical user interface, application, table

Description automatically generated

The parameter estimates show us that the regression has the equation

Y = 0.53X + 6.47

where Y represents the lung cancer death rate and X represents the cigarette sales. Based on this result, we can conclude that the slope 0.53 (95% CI: (0.36, 0.70)) signifies that for every additional unit of cigarettes sold, the rate of lung cancer death rates increases by 0.53. The intercept signifies that at 0 units of cigarettes sold, the rate of lung cancer death is 6.47 per 100 cases (95% CI: (2.15, 10.79)). The 95% confidence intervals tell us that if we were to run this study many times, 95% of all computed confidence intervals would include the true slope and interval above.

The probability of observing a slope the same as or more extreme than 0.53 is less than 0.0001. Similarly, the probability of observing an intercept the same or more extreme than 6.47 is 0.0043. Since both p-values are less than 0.05, we have sufficient evidence to reject the null hypothesis and use this equation as a regression.